Disentangled Motif-aware Graph Learning for Phrase Grounding

Zongshen Mu[1], Siliang Tang[1]*, Jie Tan[1], Qiang Yu[2], Yueting Zhuang[1]

[1]DCD Lab, College of Computer Science, Zhejiang University
[2]City Cloud Technology (China) Co., Ltd
{zongshen,siliang,tanjie95,yzhuang}@zju.edu.cn, yq@citycloud.com.cn

Abstract

In this paper, we propose a novel graph learning framework for phrase grounding in the image. Developing from the sequential to the dense graph model, existing works capture coarse-grained context but fail to distinguish the diversity of context among phrases and image regions. In contrast, we pay special attention to different motifs implied in the context of the scene graph and devise the disentangled graph network to integrate the motif-aware contextual information into representations. Besides, we adopt interventional strategies at the feature and the structure levels to consolidate and generalize representations. Finally, the cross-modal attention network is utilized to fuse intra-modal features, where each phrase can be computed similarity with regions to select the best-grounded one. We validate the efficiency of disentangled and interventional graph network (DIGN) through a series of ablation studies, and our model achieves state-of-the-art performance on Flickr30K Entities and ReferIt Game benchmarks.

Introduction

On giving an image-sentence pair, the Phrase Grounding (or more generally, Language Grounding) task aims to ground multiple noun phrases from the given sentence to the corresponding image regions, as illustrated in Figure 2. By aligning textual modality to visual modality, phrase grounding is able to link knowledge from both modalities and improves multimodal tasks that span visual question answering, visual commonsense reasoning, and robotic navigation.

Since the information that the image usually contains is more than that of the sentence, the general solution is to understand the surrounding visual context of a visual object, especially pay attention to the context considered by the sentence as well, and then map it to its referent words. Nevertheless, previous studies fail to consider such fine-grained context. Some methods neglect the importance of visual or textual context, nor do they ground individual phrases independently (Plummer et al. 2018) or sequentially (Dogan et al. 2019; Li et al. 2020). This leads to poor performance when objects with completely different functions or actions have the similar visual effect in the same image. More recently, Bajaj et al. (2019) and Liu et al. (2020) utilize fully
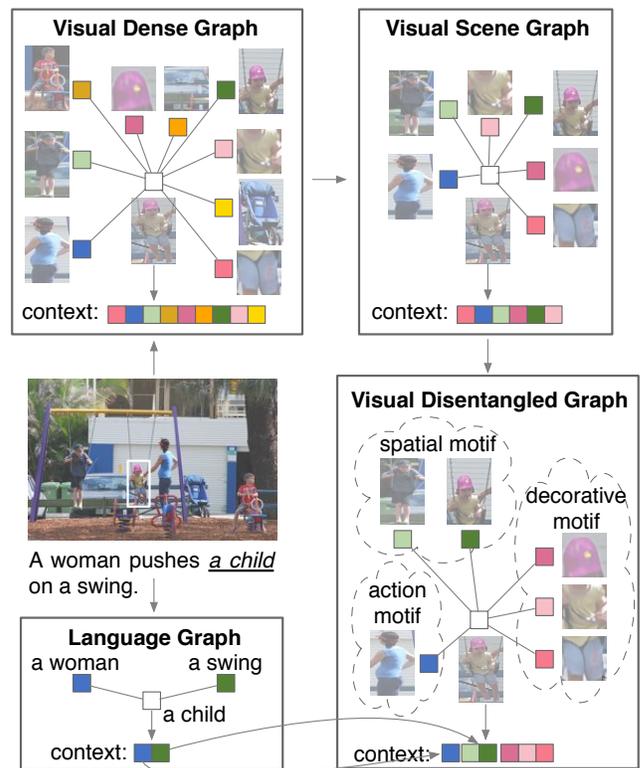

Figure 1. An example of phrase grounding.

connected dense graphs to capture coarse-grained context among visual objects, and improve grounding results.

However, their graphs have three fatal problems. First, the dense graph holds noisy false connections and makes errors in complex scenes. Second, they treat all the relations equally. Under such a framework, the node (i.e., phrase or visual object) representation is a mixture of all the possible relations. In reality, it is insufficient to reveal various high-order relation categories between nodes, or motifs[1] (e.g., *decorative*, *part-of*, *spatial*, *action*), in the contextual information. For instance, as shown in Figure 1, "a child" needs to be grounded in the image that includes multiple similar

---

[1]We post-hoc motifs referring to the method in the topic model.

object regions. As mentioned above, the language graph (the lower left of Fig. 1) is only the subgraph of visual contents. When we learn the representation of "a child" in the dense graph (the upper left of Fig. 1), it is difficult to ground the right region in noisy context. After the spurious neighbors are filtered in the scene graph (the upper right of Fig. 1), it is also difficult to make a grounding decision because too many redundant motifs around the child prevent the model from grasping the vital motif. Third, there are some co-occurring visual objects in the datasets, while some of irrelevant co-occurrences are of high probability (Wang et al. 2020), which influences the graph structure and finally damages the model's robustness and generalization capability.

In this paper, we propose the Disentangled Interventional Graph Network (DIGN), a more explainable framework for phrase grounding. First, instead of using a fully connected dense graph, we use two scene graphs to model the context among phrases and regions. Then, we use two disentangled graph networks to integrate these various context into disentangled dimensions of node (i.e., phrase or visual object) representation. Each chunk, referred to as a motif-aware embedding, can represent the strength of a certain motif (the lower right of Fig. 1), which greatly reduces the noise from spurious connections and creates fine-grained intra-modal representations for later inter-modal fusion. Furthermore, we propose the interventional visual sample synthesizing training scheme to alleviate the co-occurring biases problem and improve the model's robustness.

Our key contributions are threefold:

- We devise motif-aware graph learning based on disentangled graph network, which distinguishes diverse context and considers fine-grained contextual fusing.

- We adopt the feature and the structure interventional strategies to learn robust representations in the graph.

- We conduct a series of experiments that demonstrate our DIGN model achieves more competitive performance than the prior state-of-the-art.

Related Work

There are two lines of literature that are closely related to our work: phrase grounding and disentangling representation.

Phrase Grounidng

To ground the spatial region described by the phrase in an image, early methods (Wang et al. 2016a, 2018; Plummer et al. 2018; Akbari et al. 2019; Plummer et al. 2020) learn a shared embedding subspace and compute the distance of each phrase-region pair. The aforementioned methods ground each phrase independently and neglect textual and visual context information. One step forward, the RNN structure (Hu, Xu et al. 2016; Dogan et al. 2019) and the Transformer model (Li et al. 2020) are used to capture context among phrases and regions. To formulate more complex and non-sequential dependencies, graph architecture (Bajaj et al. 2019; Liu et al. 2020) becomes increasingly prevalent and leverage relationships between nodes to enrich representations and achieve alignment. They treat the context uni-

formly at a holistic aspect, which results in suboptimal representation and limited interpretability. An underlying fact is omitted that: grounding the phrase to the region is affected by different motifs in the graph's context. Our model disentangles context at a fine-grained level and incorporates motif-aware context into the representation.

Disentangling Representation

Learning to identify and disentangle explanatory factors hidden in the observed milieu of sensory data is the foundation of artificial intelligence. Early attempts (Tishby et al. 2015) apply the information bottleneck method only to capture and extract a relevant or meaningful data summary. Based on information-theoretic concepts, β-VAE (Higgins et al. 2017) can learn disentangled representations of independent visual data generative factors in a completely unsupervised manner. Later studies (Burgess et al. 2018; Chen et al. 2018) achieve further improvements for disentangled factor learning at a more granular level. These existing efforts are made to discover different latent factors on non-relational data. More recently, disentangled representation learning is explored in the filed of graph-structured data (Ma et al. 2019; Zheng et al. 2020), most of which mainly disentangle latent factors behind user behavior in bipartite graphs commonly seen in recommender systems. This work focuses on not only disentangling motifs in the intra-modal scene graph but also keeping disentangled in cross-modal common subspace. The disentangled graph network for phrase grounding is related to previous modular method, MattNet (Yu et al. 2018), but it is implicit to assign the meaning of the chunked representation and convenient to transfer to other tasks.

Approach

Given multiple phrases and the corresponding image, we denote a set of noun phrases as $\mathcal{P} = \{p_i\}_{i=1}^n$ and their spatial regions (or bounding boxes) as $\mathcal{B}^+ = \{b_i^+\}_{i=1}^n$ where $b_i^+ = (x_1, y_1, x_2, y_2)$ is the left-top and right-bottom coordinates. The primary goal is to select N bounding boxes from region proposals $\mathcal{B} = \{b_i\}_{i=1}^m (m >> n)$ as much overlap with ground truth annotations $\mathcal{B}^+$ as possible and get grounding results.

Figure 2 illustrates our disentangled interventional graph network (DIGN). The framework are composed of four stages to implement phrase grounding at a fine-grained level: 1) phrase and visual feature encoding; 2) two intra-modal disentangled graph networks; 3) interventional operators in the visual graph; 4) a cross-modal Transformer.

Phrase and Visual Encoders

Each phrase consists of a word or a sequence of words. For phrase features, we encode each word of the phrase to a real-valued vector by a pre-trained BERT (Devlin et al. 2019), then compute the phrase representation by taking the sum of word embeddings in each p_i. In respect of visual features, we use a region proposal network (Ren et al. 2015) to extract the top-M proposals with diversity and discrimination. Each proposal b_i is represented after the last fully-connected layer of RoI-Align pooling (He et al. 2017). To transform phrase
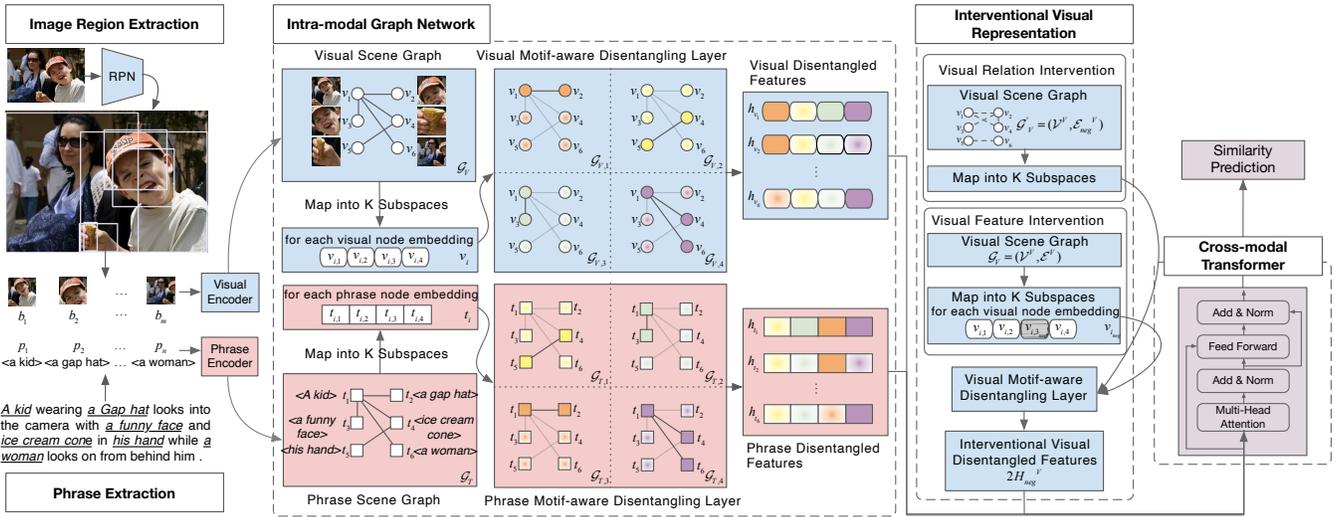
Figure 2. DIGN architecture. Node colors interpret different motifs' context, and shades of color reveal the motif's importance.

and visual features to the identical dimensions, $t_i \in \mathbb{R}^{d_{in}^t}$ and $v_i \in \mathbb{R}^{d_{in}^v}$ are obtained severally through two fully-connected layers combined with the ReLU activation function and a batch normalization layer.

Intra-Modal Graph Network

To separate various context among noun phrases and region proposals and provide explicit channels to guide motif-aware information flow, we introduce the following two intra-modal disentangled graph networks below.

Phrase Disentangled Graph Instead of a noisy dense graph to capture linguistic context, we construct a phrase scene graph $\mathcal{G}_T = (\mathcal{V}^T, \mathcal{E}^T)$ from the image caption including phrases by an off-the-shelf scene graph parser (Schuster et al. 2015). Formally, $\mathcal{V}^T = \{t_i\}_{i=1}^n$ represents nodes equipped with phrase embeddings, $(i, j) \in \mathcal{E}^T$ indicates that there is an edge connecting t_i and t_j, and $\mathcal{N}_i^T = \{t_j : (i, j) \in \mathcal{E}^T\}$ denotes the neighbors of t_i.

When grounding the phrase to the image region, the situation that similar object regions are required to be differentiated is universal and intractable. Intuitively, one phrase interacts with neighbors and makes use of its context to distinguish the regions. However, different situations need to consider the different context of the neighborhood. As shown in Figure 2, t_1 links with t_2, t_3, t_4, and t_6, whose context includes "wearing", "with", "in", and "look on from behind" high-order relation patterns. The context can be categorized into *decorative*, *part-of*, *spatial*, and *action* motifs respectively. Clearly, both the kid and the woman have funny faces, but a funny face is grounded to the kid because of his unique decorative motif (e.g., he has a gap hat) and spatial motif (e.g., the ice cream cone is on his hand), which do not belong to the context of the woman.

Therefore, unlike standard graph convolutional network (Veličković et al. 2017; Xu et al. 2019) viewing context of the ego node as a holistic one to message and update, we aim to learn a disentangled representation over various motifs, as

follows:

$$h_i^T = [h_{i,1}^T, h_{i,2}^T, ..., h_{i,K}^T] \tag{1}$$

where h_i^T is the final output of t_i, K is the hyperparameter controlling the number of latent motifs. It is expected that $h_{i,k}^T \in \mathbb{R}^{\frac{d_{out}^T}{K}}$ incorporates the contextual information of the k-th motif and is independent of other components. For the k-th chunked embeddings of all nodes, we define a motif-aware graph as $\mathcal{G}_{T,k} = \{\mathcal{V}_k^T, \mathcal{E}^T\}$. As such, we build a set of subgraphs $\mathcal{G}_T = \{\mathcal{G}_{T,1}, \mathcal{G}_{T,2}, ..., \mathcal{G}_{T,K}\}$ to refine motif-aware linguistic features through graph disentangling layers.

To separate the original mixed embedding, we project t_i into K subspaces, as shown below:

$$t_{i,k}^0 = \frac{\sigma(W_k^{T0} t_i + b_k^{T0})}{\left\| \sigma(W_k^{T0} t_i + b_k^{T0}) \right\|_2} \tag{2}$$

where $W_k^{T0} \in \mathbb{R}^{\frac{d_{out}^T}{K} \times d_{in}^T}$ and $b_k^{T0} \in \mathbb{R}^{\frac{d_{out}^T}{K}}$ are the parameters of channel k, the superscript 0 shows the original disentangled state; and $\sigma(\cdot)$ is a nonlinear activation function. We use l_2-normalization to ensure numerical stability and prevent the neighbors with overly rich features from distorting our prediction. Other components of t_i is computed as $t_{i,k}^0$, which are fed to the corresponding subgraph $\mathcal{G}_{T,k}$.

After initializing nodes of motif-aware subgraphs, we distill useful information of neighbors to comprehensively capture K latent factors' context. For each motif-aware subgraph, we design a phrase graph disentangling layer:

$$t_{i,k}^1 = \sigma(W_{ek}^{T1} t_{i,k}^0 + \sum_{j \in \mathcal{N}_i^T} a_{j,k}^{T1} W_k^{T1} t_{j,k}^0) \tag{3}$$

$$a_{j,k}^{T1} = \frac{\exp(t_{j,k}^{0\top} t_{i,k}^0)}{\sum_{k'=1}^K \exp(t_{j,k'}^{0\top} t_{i,k'}^0)} \tag{4}$$

where $a_{j,k}^{T1}$ ($a_{j,k}^{T1} \geq 0, \sum_{k'=1}^K a_{j,k}^{T1} = 1$) provide the k-th motif's importance why t_j connects with t_i; $\mathcal{N}_i$ is the first-

order neighbors of t_i; $W_{ek}^{T\,1}$ and $W_k^{T\,1} \in \mathbb{R}^{\frac{d_{out}^T}{K} \times \frac{d_{out}^T}{K}}$ represent the learnable parameters of the ego node and neighbor nodes at the k-th aspect respectively; the superscript 1 represents the output of the graph disentangling layer from one-hop neighbors; and σ is ReLU activation function followed by dropout with probability 0.5.

Equation 4 is the key step to the first-order neighbor routing. When calculating it with all neighbors, we can get the contextual information for motif k, which is similar to the attention mechanism, but not exactly the same. The attention mechanism like Graph Attentive networks (Wang et al. 2019) coarsely aggregates one-hop neighbors and updates the ego node as an entirety, causing difficult to distinguish different context and entangling latent motifs. Conversely, the neighbor rooting mechanism slices the node into k components, for which each part can attend similar context and provide an interpretable result.

Having used one-hop neighbors, we can stack more graph disentangling layers to gather influential signals from higher-order neighbors. After L layers, we sum up the motif-aware representations at different layer as the k-th chunk representation:

$$h_{i,k}^T = t_{i,k}^0 + t_{i,k}^1 + ... + t_{i,k}^L \tag{5}$$

We argue that $h_{i,k}^T$ describes the semantic context related with the k-th motif. Finally, we joint K chunks to update t_i after L disentangling layers, as presented in equation 1. Consequently, we not only disentangle phrase representations but also explain each part of them.

Visual Disentangled Graph We establish a visual scene graph $\mathcal{G}_V = (\mathcal{V}^V, \mathcal{E}^V)$ by the scene graph generation model (Zellers et al. 2018) that exploits visual relationships among region proposals.

We hypothesize that the visual graph context implies the same categories of motifs as those in the context of the phrase graph. So we also slice the visual node embedding into K components, and every chunked node embeddings constructs independent subgraph $\mathcal{G}_{V,k} = (\mathcal{V}_k^V, \mathcal{E}^V)$. $\mathcal{G}_V$ is composed of K motif-aware graphs $\{\mathcal{G}_{V,1}, \mathcal{G}_{V,2}, ..., \mathcal{G}_{V,K}\}$. Similar to the phrase graph network, we use another disentangled graph network to integrate visual context containing varying motifs into representations. More details are presented in Algorithm 1 of Appendix A.

We firstly map each v_i into K subspaces to initialize K chunked embeddings by linear transformations:

$$v_{i,k}^0 = \frac{\sigma(W_k^{V\,0} v_i + b_k^{V\,0})}{\left\| \sigma(W_k^{V\,0} v_i + b_k^{V\,0}) \right\|_2} \tag{6}$$

where $W_k^{V\,0} \in \mathbb{R}^{\frac{d_{out}^V}{K} \times d_{in}^V}$ and $b_k^{V\,0} \in \mathbb{R}^{\frac{d_{out}^V}{K}}$, $v_{i,k}^0$ is the k-th part of v_i^0. The aggregation and updating steps of a visual graph disentangling layer about the k-th motif can be described as:

$$v_{i,k}^1 = \sigma(W_{ek}^{V\,1} v_{i,k}^0 + \sum_{j \in \mathcal{N}_i^V} a_{j,k}^{V\,1} W_k^{V\,1} v_{j,k}^0) \tag{7}$$

where $W_{ek}^{V\,1}$, $W_k^{V\,1} \in \mathbb{R}^{\frac{d_{out}^V}{K} \times \frac{d_{out}^V}{K}}$, $a_{j,k}^{V\,1}$ is the result of the neighbor rooting mechanism in visual graph network. After

iterating more layers, the final output of each chunk $h_{i,k}^V$ is obtained by summing up every layers' hidden states of v_i:

$$h_{i,k}^V = v_{i,k}^0 + v_{i,k}^1 + ... + v_{i,k}^L \tag{8}$$

The disentangled result h_i^V is depicted to:

$$h_i^V = [h_{i,1}^V, h_{i,2}^V, ..., h_{i,K}^V] \tag{9}$$

Interventional Visual Representation

The dataset co-occurring bias, the irrelevant relations, may lead to in-correct attention. When modeling the context of regions in the image by disentangled graph network, it is possible that the biased data plagues chunked representation with irrelevant information through context aggregation. Notably, visual features from the encoder are vulnerable to learn the bias (Cadene et al. 2019). Thus, we intervene in representation learning in disentangled graph network to reduce the effect of bias and robust representations.

The core idea of intervention is to change the environment, helping the model train with more unseen data and discover more possible reasons. We utilize the interventional strategies to improve motif-aware learning during the training process. It is efficient and effective to construct interventional samples at the feature and the structure levels in the graph-structure data. Compared to the masked attentive region strategy (Liu et al. 2019), our strategies do not depend on language attention and are more flexible to learn context-aware features in the graph. The pseudocode of the interventional process is listed in Appendix B.

We first introduce the strategy of structure intervention. We aim to change the original edges in the visual scene graph. When modifying the nodes of edges, we just change the target node of the directional edge. In this way, it can eliminate the impact of changing other connected nodes' features as much as possible. We interrupt and randomly interchange all the target nodes' edges in the visual graph. Integrated unrelated motif context into the representation by the neighbor routing mechanism, these contrastive samples help the network to grasp natural motifs between nodes.

The feature interventional samples are synthesized with the implementation of masking right chunked features with incorrect data. There are two types of masked feature interventional methods. One type is to replace features of neighbor nodes in corresponding dims. Neighbors are usually analogous to the ego node on the semantic aspect, making the model distinguish similar cases. The other type is to fill each dim of chunked embedding with noisy distribution. The generated features are treated as noisy training data. It is beneficial and general for the model to learn more unseen samples and have the resilience to noisy data. Therefore, the interventional samples can promote the model to find and understand the meaning of the replaced chunk automatically for phrase grounding.

We define the particular loss function in Section to make full use of interventional samples. Moreover, the intervention module is vital to motif-aware learning, which is demonstrated in the following ablation studies.

Cross-Modal Transformer

To align two modal representations in a shared space, we introduce a cross-modal Transformer, as shown in the last column of Figure 2. Unlike simple concatenation of two modal features, the multi-head attention mechanism can benefit each head block to capture the corresponding motif's mutual information from both modal disentangled representations. When the visual information is injected to the linguistic encoding and the linguistic information is incorporated to the visual encoding, the Transformer can dynamically select judicious cues for target representation.

We denote the inputs for the Transformer as $h^T = \{h_1^T, h_2^T, ..., h_n^T\}$ and $h^V = \{h_1^V, h_2^V, ..., h_m^V\}$, which are processed by phrase and visual graph networks respectively. Specifically, multi-head attention is computed as Multihead(Q, K, V), where Q is the query, K is the key, and V is the value. The details can refer to Vaswani et al. (2017). Two modal features are interacted by:

$$c^T = \text{Multihead}(W_q h^T, W_k h^V, W_v h^V) \tag{10}$$

$$c^V = \text{Multihead}(W_q h^V, W_k h^T, W_v h^T) \tag{11}$$

where W_q, W_k and W_v are learnable weights, c^T is the blended feature from phrases, while c^V is the guided representation from regions.

Through this form Transformer, linguistic and visual features are associated so that the similarity prediction is defined as:

$$\text{sim}(c_i^T, c_j^V) = c_i^T \cdot c_j^V \tag{12}$$

where we just use an inner (dot) product to achieve state-of-the-art performance. The similarity score is thought as the the probability whether the phrase p_i is grounded to the image region b_i.

Training Objectives

Our loss function includes three parts. The first part is the phrase independent loss. For encouraging K components of the node embedding to be independent, we apply distance correlation (Székely et al. 2007) to characterizing independence of any two paired vectors. Each pair's coefficient is zero if and only if these vectors are independent. After phrase disentangled graph network, we formulate the independent loss of phrase chunked embeddings as:

$$\ell_{ind}^T = \sum_{i=1}^n \sum_{k=1}^K \sum_{k'=k+1}^K \rho(h_{i,k}^T, h_{i,k'}^T) \tag{13}$$

$$\rho(h_{i,k}^T, h_{i,k'}^T) = \frac{Cov(h_{i,k}^T, h_{i,k'}^T)}{\sqrt{D(h_{i,k}^T) \cdot D(h_{i,k'}^T)}} \tag{14}$$

where $Cov(\cdot)$ and $D(\cdot)$ represent the covariance and the variance between vectors respectively. Identically, the visual independent loss as the second part is computed as:

$$\ell_{ind}^V = \sum_{i=1}^m \sum_{k=1}^K \sum_{k'=k+1}^K \rho(h_{i,k}^V, h_{i,k'}^V) \tag{15}$$

The third part is cross-modal grounding loss. For a phrase p_i, the area of best-matched result b_i in $\mathcal{B}$ is maximum, overlapping with the ground truth b_i^+. We denote $\{\mathcal{B} - b_i\}$ as

$M - 1$ failed results. For outputs of the interventional model, we change their representations and treat them unmatched with ground truth. In total, there are $2M - 1$ negative samples. Instead of minimizing the negative log-likelihood of correct correspondence scores, we consider a contrastive loss function, called InfoNCE (Oord, Li, and Vinyals 2018), in this paper:

$$\ell_{ground} = \sum_{i=1}^n -\log \frac{\exp(\text{sim}(p_i, b_i)/\tau)}{\sum_{j=1}^{2M-1} \exp(\text{sim}(p_i, b_j)/\tau)} \tag{16}$$

where $\text{sim}(\cdot)$ is the output of similarity prediction, the sum is over one positive and $2M - 1$ negative regions, τ denotes the temperature parameter. During training process, for any phrase p_i, the model is tuned to maximize the numerator of the log argument and minimize its denominator as well. As a result, the model can thoroughly learn the difference between the true one and interventional samples.

Overall, all the losses are optimized to update learnable parameters:

$$\mathcal{L} = \ell_{ind}^T + \ell_{ind}^V + \ell_{ground} \tag{17}$$

Experiments and Results

Datasets and Evaluation

We validate our model on two common datasets for phrase grounding.

Flickr30K Entities (Plummer et al. 2015) contains 31,783 images where each image corresponds to five captions with annotated noun phrases. Consistent with the work (Dogan et al. 2019), when a single phrase is annotated with multiple ground-truth bounding-box, we merge the boxes and use the union region as their ground-truth. We divide the dataset into 30k images for training, 1k for validation, and 1k for testing.

ReferIt Game (Kazemzadeh et al. 2014) contains 20,000 images along with 99,535 segmented image regions. Each image is equipped with multiple referring phrases and corresponding bounding-boxes. We use the same split as Akbari et al. (2019), which contains 10k training and 10k test images. And we add the supplementary experiment on the Ref-COCO+ in Appendix D.

A noun phrase in the dataset is grounded correctly if and only if the predicted box and its ground-truth have at least 0.5 IoU (intersection over union). Based on the criteria, our measure of performance is grounding accuracy, which is the ratio of correctly grounded noun phrases to the total number of phrases in the test set.

Experimental Setup

We use the uncased Bert[1] to get 768-dim phrase embeddings and a java toolkit[2] to parse phrase scene graph for each caption. For visual scene graph generation[3], we use Faster R-CNN with VGG-16 backbone as a mechanism to extract top-100 proposal regions with 2048-dimension features and top-500 relations. All the dimensions of d_{in}^T, d_{in}^V,

---

[1]https://pypi.org/project/pytorch-pretrained-bert/

[2]https://nlp.stanford.edu/software/scenegraph-parser.shtml

[3]https://github.com/rowanz/neural-motifs

Model Type	Method	Flickr30k	ReferIt
Independent	SMPL (Wang et al. 2016b)	42.08	-
	NonlinearSP (Wang et al. 2016a)	43.89	-
	SCRC (Hu, Xu et al. 2016)	-	17.93
	GroundeR (Rohrbach et al. 2016)	47.81	26.93
	MCB (Fukui, Park et al. 2016)	48.69	-
	RtP (Plummer et al. 2015)	50.89	-
	SimNet (Wang et al. 2018)	51.05	31.26
	CGRE (Luo et al. 2017)	-	31.85
	MSRC (Chen et al. 2017a)	57.53	32.21
	IGOP (Yeh et al. 2017)	53.97	34.70
	SPC+PPC (Plummer et al. 2015)	55.49	-
	CITE (Plummer et al. 2018)	59.27	34.15
	MultiG (Akbari et al. 2019)	69.19	62.76
	ZsgNet (Sadhu et al. 2019)	63.39	58.63
	OneStage (Yang et al. 2019)	68.69	59.30
	SSNs (Plummer et al. 2020)	72.50	60.50
Sequential	QRC Net (Chen et al. 2017b)	65.14	44.07
	SeqGROUND (Dogan et al. 2019)	61.60	-
Graph	GG++ (Bajaj et al. 2019)	66.93	44.91
	VisualBERT (Li et al. 2020)	71.33	-
	LCMCG (Liu et al. 2020)	76.74	-
	DIGN	**78.73**	**65.15**

Table 1. Overall accuracy(%) comparison on two datasets

d_{out}^T and d_{out}^V are set as 512. For phrase and visual disentangled graph networks, the number of neighbor rooting layer is 2. We do a series of experiments to study the impact of the chunked number K in $\{1, 2, 4, 8, 16\}$ and evaluate on two datasets. The results on Flickr30K is 77.44%, 78.26%, 78.73%. 78.56% and 78.08% respectively. We can see our model reaches the best performance at 4. The accuracy on ReferIt is 63.94%, 64.34%, 65.15%, 64.14% and 64.83% respectively. It is observed that the accuracy fluctuates after the k set to 4. Considering effectiveness and efficiency, we set K to 4. We only use one layer of Transformer with 4 multiheads as cross-modal mapping. For the InfoNCE loss, the hyperparameter τ is set to 0.2. We train the end-to-end network by the SGD optimizer with learning rate 1e-3, weight decay 1e-4, and momentum 0.9. The model iterates 6 epochs with batch size 32.

Results and Comparison

The phrase grounding results of two benchmarks are reported in Table . We classify prior works and ours into three classes (i.e. independent, sequential and graph models) discussed in Section . Our model surpasses all state-of-the-art techniques and achieves the best performances of 78.73% and 65.15% on Flickr30K and ReferIt respectively. It can be apparently seen that our method achieves an absolute increase of ∼2% compared to LCMCG on Flickr30K. Although both methods consider the context, our model learns the disentangled motif-aware context in scene graphs, which

Models	Flickr30K	ReferIt
MLPs + MLPs Fusion	65.07	35.46
GraphAttentionNetwork + MLPs Fusion	72.44	60.31
DisentangledGraphNetwork + MLPs Fusion	74.34	61.71
DGN + Cross-Modal Transformer	75.15	62.74
DGN + CMT + Structure Intervention	77.44	63.87
DGN + CMT + Feature Intervention	76.86	64.35
DIGN	**78.73**	**65.15**

Table 2. Ablation results on Flickr30k and ReferIt datasets.

is the reason why our model can be improved. Comparing to MultiGrounding on ReferIt, we argue that the disentangled visual context and interventional sample strategies are mainly gain of improvement about ∼2.4%. The fine categories comparison is analyzed in Appendix C.

Ablation Studies

Extensive experiments are conducted to investigate the benefits of each block, with results enclosed in Table . To be more specific, the effectiveness of the disentangled graph network, two kinds of interventional strategies, and the cross-modal Transformer are clarified as follows.

The first line is used to validate the performance of our model is from our model design rather than better pre-trained extractor. The baseline (the second line) utilizes the graph attention mechanism (GAT) in scene graphs and MLPs in inter-modal fusion for grounding. Disentangled graph network (DGN) is substituted for GAT to learn disentangled embeddings, which boosts the performance from 72.44% to 74.34% and from 60.31% to 61.71% on Flickr30K and ReferIt respectively. The third line replaces MLPs with a cross-modal Transformer (CMT) to transform K independent components of intra-modal embedding to common subspaces and achieve alignment, improving accuracy within 0.8%∼1%. The fourth and fifth lines are added to the third one by intervening in structure and feature levels respectively. Against the third line, the fourth one enhances the performance from 75.15% to 77.44% on Flickr30K and from 62.74% to 63.87% on ReferIt. In contrast to the third one, the fifth line has 1.7% and 1.6% impact on two benchmarks. Particularly, the performance of the structure intervention on Flickr30K is more influential than that of in the performance than the feature intervention. Because the number of images on Flickr30K is more than that on ReferIt, it is reasonable that the model on Flickr30K grasps more rich and complex context. The last line is our whole model and obtains the highest score across datasets. The main improvement can be shown from the second line to the second line (1.4%∼1.9%) and from the third line to the last line (2.4%∼3.6%). It is strongly proved that DGN and interventional methods are the most critical modules. In summary, The results show consistent patterns of effects in two datasets, showcasing that our model can generalize for the phrase grounding tasks.
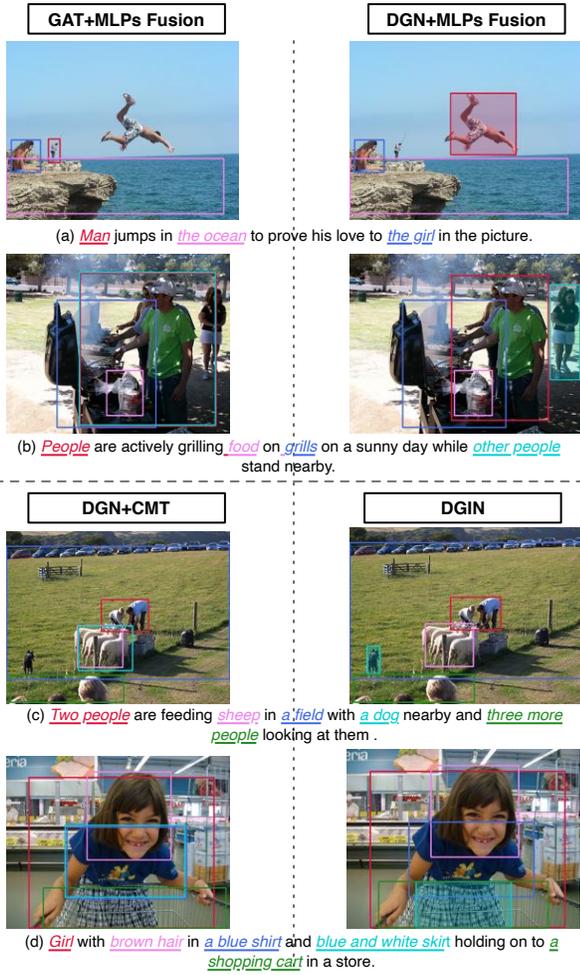
Figure 3. The visualization examples of models.

Qualitative Results

From Figure 3, we can see the contributions of the disentangled graph network module and interventional module. The upper part of the figure compares the graph attention network with the disentangled graph network, which proves the disentangled module can discriminate similar objects by chunked representations rather than messed representations. The lower part provides some interesting data, in terms of the fact that the interventional module can reduce the bias and break spurious relations at the structure level (e.g. the dog may usually co-occur with people and streets instead sheep) and recognize slight differences between objects at the feature level (e.g. the features of "blue shirt" and "blue and white skirt" are replaced with others to learn intrinsic color character).

We empirically inspect what kind of motifs are picked up by visualizing the weights of neighborhood routing. Figure 4 shows the post-hoc explanations of learned spatial motif and part-of motif. There are more details in Appendix D. The left column shows two latent motifs of context in the phrase scene graph, and the right column exhibits motifs corresponding to phrase ones in the visual scene graph.
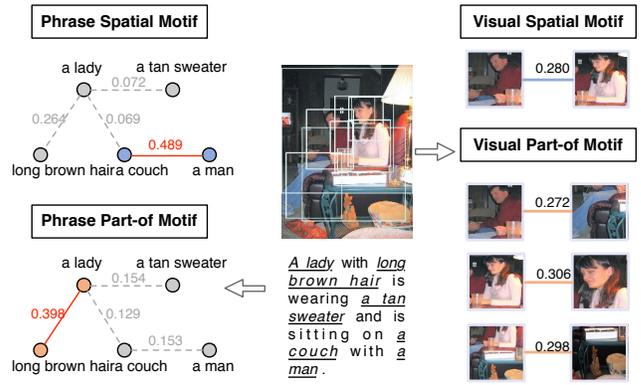

Figure 4. Illustration of motif-aware graph learning.

For the phrase subgraph, the colored nodes represent that they are the most related in the current motif against other motifs. There are too many relationships in the visual scene graph, which are difficult to show a complete graph in the figure. So we just visualize the most prominent motif context between chunked nodes by the colored edges. From the results in Figure 4, the motif-aware context incorporated by the visual graph is more than that in the phrase graph even though disentangled. Hence, it is reasonable to compute the similarity of corresponding chunked embeddings to achieve grounding. Additional grounding results of our model are provided in Appendix E.

Conclusion

In this paper, we propose a disentangled interventional motif-aware learning framework for the phrase grounding task. Disentangled graph network distinguishes the importance of different motifs, integrates motif-aware context into node representations in intra-modal scene graphs, and provides interpretable results. Then interventional schemes at the aspects of feature and structure improve the resilience to the biased data and the ability of generalization. Finally, fine-grained intra-modal representations are fused in the cross-modal Transformer to finish phrase grounding. Our method (DGIN) is demonstrated on two public datasets, outperforming state-of-the-art by a considerable increase. In our future work, these tasks, constructing a more accurate scene graph and grounding phrase, should be considered to enhance each other. Additionally, we would like to make in-depth analyses of cross-modal fusing.

Acknowledgments

References

Akbari, H.; et al. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *CVPR*.

Bajaj, M.; et al. 2019. G3raphGround: Graph-Based Language Grounding. In *ICCV*.

Burgess, C. P.; et al. 2018. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599* .

Cadene, R.; et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*.

Chen, K.; Kovvuri, R.; Gao, J.; and Nevatia, R. 2017a. MSRC: Multimodal spatial regression with semantic context for phrase grounding. In *ICMR*.

Chen, K.; et al. 2017b. Query-guided regression network with context policy for phrase grounding. In *ICCV*.

Chen, R. T.; et al. 2018. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Dogan, P.; et al. 2019. Neural sequential phrase grounding (seqground). In *ICCV*.

Fukui, A.; Park, D. H.; et al. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

Higgins, I.; et al. 2017. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Hu, R.; Xu, H.; et al. 2016. Natural language object retrieval. In *CVPR*.

Kazemzadeh, S.; et al. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2020. What Does BERT with Vision Look At? In *ACL*.

Liu, X.; et al. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*.

Liu, Y.; Wan, B.; Zhu, X.; and He, X. 2020. Learning Cross-Modal Context Graph for Visual Grounding. In *AAAI*.

Luo, R.; et al. 2017. Comprehension-guided referring expressions. In *CVPR*.

Ma, J.; Cui, P.; Kuang, K.; Wang, X.; and Zhu, W. 2019. Disentangled graph convolutional networks. In *ICML*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .

Plummer, B. A.; Kordas, P.; Hadi Kiapour, M.; Zheng, S.; Piramuthu, R.; and Lazebnik, S. 2018. Conditional image-text embedding networks. In *ECCV*.

Plummer, B. A.; et al. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Plummer, B. A.; et al. 2020. Shapeshifter Networks: Cross-layer Parameter Sharing for Scalable and Effective Deep Learning. *arXiv preprint arXiv:2006.10598* .

Ren, S.; et al. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.

Sadhu, A.; et al. 2019. Zero-shot grounding of objects from natural language queries. In *ICCV*.

Schuster, S.; et al. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 70–80.

Székely, G. J.; Rizzo, M. L.; Bakirov, N. K.; et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35(6): 2769–2794.

Tishby, N.; et al. 2015. Deep learning and the information bottleneck principle. In *ITW*, 1–5.

Vaswani, A.; et al. 2017. Attention is all you need. In *NeurIPS*.

Veličković, P.; et al. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* .

Wang, L.; et al. 2016a. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Wang, L.; et al. 2018. Learning two-branch neural networks for image-text matching tasks. *TPAMI* 41(2): 394–407.

Wang, M.; et al. 2016b. Structured matching for phrase localization. In *ECCV*.

Wang, P.; et al. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual commonsense r-cnn. In *CVPR*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *ICLR*.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *ICCV*.

Yeh, R.; et al. 2017. Interpretable and globally optimal prediction for textual grounding using image concepts. In *NeurIPS*.

Yu, L.; et al. 2018. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*.

Zellers, R.; et al. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*.

Zheng, Y.; Gao, C.; Li, X.; He, X.; Li, Y.; and Jin, D. 2020. Disentangling User Interest and Popularity Bias for Recommendation with Causal Embedding. *arXiv preprint arXiv:2006.11011* .